JP1265378A MicroPatent Report

EUROPEAN CHARACTER RECOGNIZING SYSTEM

[71] Applicant: FUJITSU LTD

[72] Inventors: SATO JUN

[21] Application No.: JP198893819A

[22] Filed: 19880415

[43] Published: 19891023

[30] Priority: JP JP198893819A 19880415

[No drawing]

Go to Fulltext Get PDF

[57] Abstract:

PURPOSE: To correctly recognize European characters in a short time even in case the space between characters is small and the characters are in contact with each other by calculating the distance between the geometrical features of an extracted word and those of each word registered in a dictionary and deciding the words most coincident with each other.
CONSTITUTION: The words are segmented by a word segmenting part 10 out of character lines supplied by the input of an image obtained by reading optically an European language document. Then a feature extracting part 12 extracts the geometrical features of the segmented word. The vertical line distribution in a word, the vertical line density distribution in a word and/or the loop part distribution in a word are registered in a dictionary 14 for each word as the word features. Then the difference between the geometrical features of the word extracted at the part 12 and the features of each word registered in the dictionary 14 is decided. A deciding part 16 decides the words most coincident with each other. Thus the European characters can be correctly recognized in a short time even in case the space between characters is small and the characters are in contact with each other.
COPYRIGHT: (C)1989,JPO&Japio

[52] US Class:

[51] Int'l Class: G06K000972

[52] ECLA:

⑲日本国特許庁（JP）　⑪特許出願公開

⑫公開特許公報（A）　平1－265378

⑤Int. Cl.⁴	識別記号	庁内整理番号	㊸公開　平成1年(1989)10月23日
G 06 K 9/72		6942－5B	

審査請求　未請求　請求項の数 2　（全5頁）

㊹発明の名称　欧文文字認識方式

㉑特　願　昭63－93819
㉒出　願　昭63(1988) 4月15日

㉒発明者　佐藤　純　神奈川県川崎市中原区上小田中1015番地　富士通株式会社内
㉑出願人　富士通株式会社　神奈川県川崎市中原区上小田中1015番地
㉔代理人　弁理士　井桁　貞一　外2名

明細書

1．発明の名称

欧文文字認識方式

2．特許請求の範囲

（1）欧文文書を光学的に読取って認識する欧文文字認識方式に於いて、

欧文文字行の中から単語を切り出す単語切出し部（10）と；

該単語切出し部（10）で切り出された単語の幾何学的特徴を抽出する特徴抽出部（12）と；

各単語の幾何学的特徴を予め登録した辞書（14）と；

前記特徴抽出部（12）で抽出された単語の幾何学的特徴と前記辞書（14）に登録された各単語の特徴との距離を演算して最も合致する単語を判定する判定部（16）と；

を備えたことを特徴とする欧文文字認識方式。

（2）前記単語の幾何学的特徴として、単語内の縦線分布、単語内の縦方向線密度分布及び又は単語内のループ部分の分布を用いることを特徴とする請求項1記載の欧文文字認識方式。

3．発明の詳細な説明

［概要］

欧文文書を光学的に読取って認識する欧文文字認識方式に関し、

文字間隔が狭く文字同志が接触している場合にも、短時間の処理で正しい認識結果が得られることを目的とし、

欧文文字行の中から単語を切出して単語単位で幾何学的特徴、即ち、単語内の縦線分布、単語内の縦方向線密度分布、及び又は単語内のループ部分の分布を抽出し、抽出した単語の幾何学的特徴と辞書に登録された各単語の幾何学的特徴との距離を演算して最も合致する単語を判定するように構成する。

［産業上の利用分野］

本発明は、欧文文書を光学的に読取って認識す

る欧文文字認識方式に関する。

文字読取装置における文字認識方式にあっては、光学的に読取った文書中の文字の領域を１個ずつ決定して文字切出しを行なって上で文字の特徴を抽出し、辞書に登録された各文字の特徴との距離を演算して最も合致する文字を判定しており、文字認識率を向上させることが望まれる。

［従来の技術］

従来の欧文文字の認識方式にあっては、光学的に読取った欧文文書中の文字領域を１個ずつ決定して文字を切出し、切出された文字単位で辞書との比較（距離演算）により文字を認識している。

［発明が解決しようとする課題］

しかしながら、文書中の文字領域を１個ずつ決定した上で文字を切出して認識する従来方式にあっては、文字間隔が狭いことによって隣接する文字同志が接触している場合等には、正常に文字切出しが行なわれず、正しい文字認識結果が得られない場合がある。

また、文字切出しの誤りを修正するために、複数の切出し候補について文字認識をおこなう方式や、切出し位置を変化させながら文字認識を行ない、適切な認識結果が得られるまで処理を繰り返す方式等が試みられている

しかし、これらの方式は試行回数が増大するために処理時間が長くなるという問題がある。

特に文字間隔が狭い場合の欧文文字の文字切出しの困難さは、各文字部分の切出し範囲を幾何学的な条件のみにより推定していることに起因している。

例えば日本語の活字認識においては、「文字は略正方形であり、且つ各文字の幅は略一定である。」という幾何学的条件を用いて各文字範囲を推定することが可能であるが、欧文の場合には、文字の種類によって文字幅が変化するため、このような単純な条件は使用できない。例えば、「ｍ」は「ｉ」の２倍以上の文字幅をもっている。

このため欧文の個々の文字範囲を推定するためには、幾何学的条件のみならず、文字としての認識結果を利用して「文字としての妥当性」を確認して各文字の範囲を決定する必要がある。

しかし、「文字としての妥当性」を判定するだけでは不十分な場合がある。例えば「rn」という文字画像は、２つに分割して「ｒ」＋「ｎ」とも認識可能であるし、１つに統合して「ｍ」とも認識可能である。「ｒ」＋「ｎ」か「ｍ」かは意味判断を伴わずに判定することは困難且つ不確実であり、この結果、欧文の文字認識をより一層困難なものにしている。

本発明は、このような従来の問題点に鑑みてなされたもので、文字間隔が狭く文字同志が接触している場合にも、短時間の処理で正しい認識結果が得られる欧文文字認識方式を提供することを目的とする。

［課題を解決するための手段］

第１図は本発明の原理説明図である。

第１図において、欧文文書を光学的に読取った画像入力による文字行の中から単語切出し部１０によって単語を切出し、特徴抽出部１２により切出した単語の幾何学的特徴を特徴を抽出する。単語の幾何学的特徴としては、例えば、単語内の縦線分布、単語内の縦方向線密度分布、及び又は点後内のループ部分の分布を抽出する。

更に、各単語の幾何学的特徴を予め登録した辞書１４が設けられる。辞書１４にも各単語毎に単語内の縦線分布、単語内の縦方向線密度分布、及び又は単語内のループ部分の分布が単語の特徴として登録されている。

そして、特徴抽出部１２で抽出された単語の幾何学的特徴と辞書１４に登録された各単語の特徴との距離を演算して最も合致する単語を判定部１６により判定する。

［作用］

このような本発明の欧文文字認識方式にあっては、欧文文書のもつ特徴として「分かち書きにより単語単位で分割されている」点に着目し、単語

単位に幾何学的条件、即ち、単語内の縦線分布、単語内の縦方向線密度分布、単語内のループ部分の分布等を判定しつつ単語としての意味判定を行なうことにより、確実に欧文文書を認識することができる。

また文字切出しは単語単位で行なうことから、文字間隔が狭い場合であっても、単語単位の分かち書きによって単語間のスペースから確実に単語単位の切出しができ、文字単位の切出しのような困難さは解消され、更に単語単位で切出して認識することから文字単位の切出し認識に比べ認識処理時間も大幅に短縮できる。

［実施例］

第２図は本発明の一実施例を示した実施例構成図である。

第２図において、１８は画像入力部であり、欧文文書を光学的に読取り、光学的な読取りで得られたアナログ画像信号を２値画像データに変換して画像メモリ２０に格納する。２２は行抽出部であり、画像メモリ２０に格納された欧文文書の画像データの中から行毎の画像データを抽出して行画像メモリ２４に１行分の画像データを記憶する。

２６は縦投影作成部であり、行画像メモリ２４に蓄積された１行文の画像データを読出して縦投影データを作成する。縦投影作成部２６で作成された１行文の画像データの縦投影データは縦投影判別部２８に与えられ、縦投影判別部２８で単語間の空白を検出して単語の範囲を決定する。縦投影判別部２８で単語の範囲が決定されると、この決定情報を受けて単語切出し部１０が１行文の画像データの中から単語データを切出して単語画像メモリ３０に格納する。

単語画像メモリ３０に１行文の各単語データが格納されると、特徴抽出部１２が起動し単語画像メモリ３０から１つづつ単語画像データを取り込んで単語の幾何学的特徴を抽出する。

特徴抽出部１２で抽出される単語の幾何学的特徴としては、

（ａ）単語内の縦線分布

（ｂ）単語内の縦方向線密度分布

（ｃ）単語内のループ部分の分布

を単語の幾何学的特徴として抽出する。この実施例にあっては、単語内の縦線分布と単語内の縦方向線密度分布の２つを幾何学的特徴として抽出している。更に単語の幾何学的特徴として、例えば単語の上凸カーブの分布や下凸カーブの分布等を抽出してもよい。

一方、１４は辞書であり、欧文の各単語毎に特徴抽出部１２で抽出する単語の幾何学的特徴と同じ特徴を予め抽出した結果が各単語単位で登録されている。

１６は判定部であり、距離計算部３２と単語判定部３４を備える。距離計算部３２は特徴抽出部１２より抽出された単語の幾何学的特徴の入力を受けたときに、辞書１４に予め登録されている各単語の幾何学的特徴との間の距離（相違度）を演算する。単語判定部３４は距離計算部３２の各単語毎の計算距離を受けて最も計算距離の小さい単語を認識結果として判定する。

次に第３図の認識処理説明図を参照して第２図の実施例の動作を説明する。

今、行抽出部２２により行画像メモリ２４に格納された１行文の画像データの中から単語切出し部１０により単語画像メモリ３０に第３図に示す「ｃｏｍｍｕｎｉｃａｔｉｏｎ」が切出されたとする。この単語画像メモリ３０の単語について、特徴抽出部１２は第３図に示すように縦線分布としての縦線特徴、及び縦方向線密度としての線密度特徴のそれぞれを抽出する。すなわち、縦線特徴とは入力単語「ｃｏｍｍｕｎｉｃａｔｉｏｎ」の単語内における縦方向の線分布であり、一方、線密度特徴とは入力単語「ｃｏｍｍｕｎｉｃａｔｉｏｎ」における横方向の線密度を加算したデータである。例えば「ｃ」を例にとると、横方向に２本の線成分が存在することから「凸」状の線密度特徴が抽出される。

一方、辞書１４には例えば「ｃｏｍｍｕｎｉｃａｔｅ」に対応した縦線特徴及び線密度特徴のそれぞれが登録されている。

その結果、距離計算部３２では特徴抽出部１２で抽出された入力単語「ｃｏｍｍｕｎｉｃａｔｉｏｎ」の縦線特徴及び線密度特徴のそれぞれにつき、辞書１４に登録された辞書単語「ｃｏｍｍｕｎｉｃａｔｅ」の縦線特徴及び線密度特徴との間の距離を計算し、この距離の計算結果から単語判定部３４において入力単語「ｃｏｍｍｕｎｉｃａｔｉｏｎ」に対し距離の最も小さい辞書単語が「ｃｏｍｍｕｎｉｃａｔｅ」であることを判定し、更に入力単語の語尾「ｉｏｎ」と辞書単語の語尾「ｅ」の相違から入力単語が名詞形「ｃｏｍｍｕｎｉｃａｔｉｏｎ」であることを最終的に判定して入力単語を認識することができる。

尚、上記の実施例は単語の幾何学的特徴として縦線分布及び縦方向線密度の分布の２つを用いた場合を例にとるものであったが、これに加えて単語内のループ部分の分布等を特徴として抽出するようにしてもよい。

［発明の効果］

以上説明してきたように本発明によれば、欧文文字の認識において文字フォント（字形）や文字サイズが異なったり、文字間隔に広狭があっても単語単位に得られる所定の幾何学的特徴を抽出することにより文字同志が接触している場合にあっても正確に文字を認識して読取ることができる。

また、単語単位で１回だけ特徴抽出及び辞書検索を行なう方式であるため、従来の文字単位での特徴抽出及び辞書検索する方式に比べ、認識処理を高速化することができる。例えば、抽出する特徴次元数を同一にすると従来方式に比べ、本発明にあっては平均で５倍程度高速の処理を実現することができる。

４．図面の簡単な説明

第１図は本発明の原理説明図；

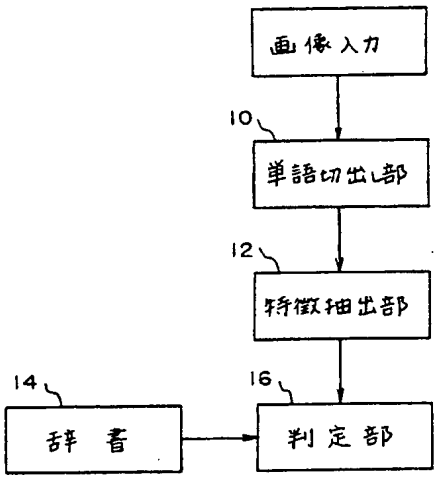
第２図は本発明の実施例構成図；

第３図は本発明の認識処理説明図である。

図中、

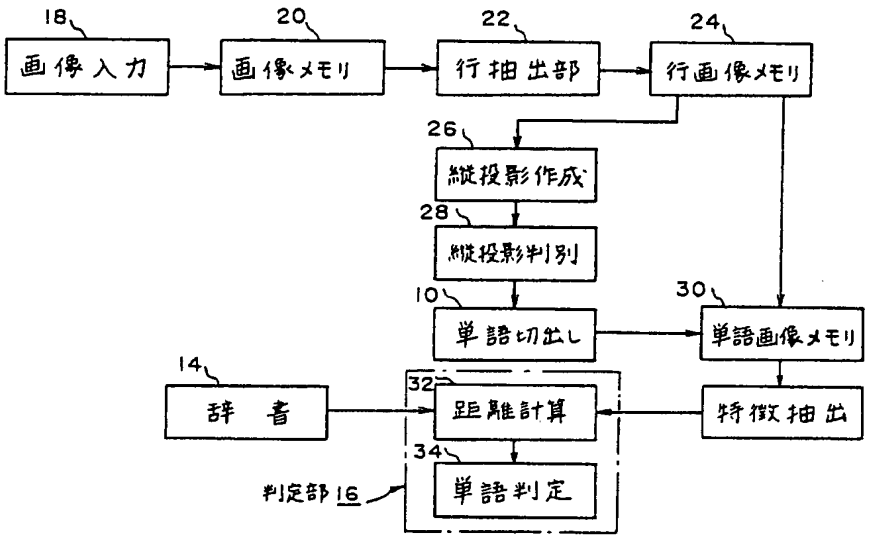
１０：単語切出し部
１２：特徴抽出部
１４：辞書
１６：判定部
１８：画像入力部
２０：画像メモリ
２２：行抽出部
２４：行画像メモリ
２６：縦投影部
２８：縦投影判別部
３０：単語画像メモリ
３２：距離計算部
３４：単語判定部

特許出願人　富士通株式会社
代理人　弁理士　井桁貞一



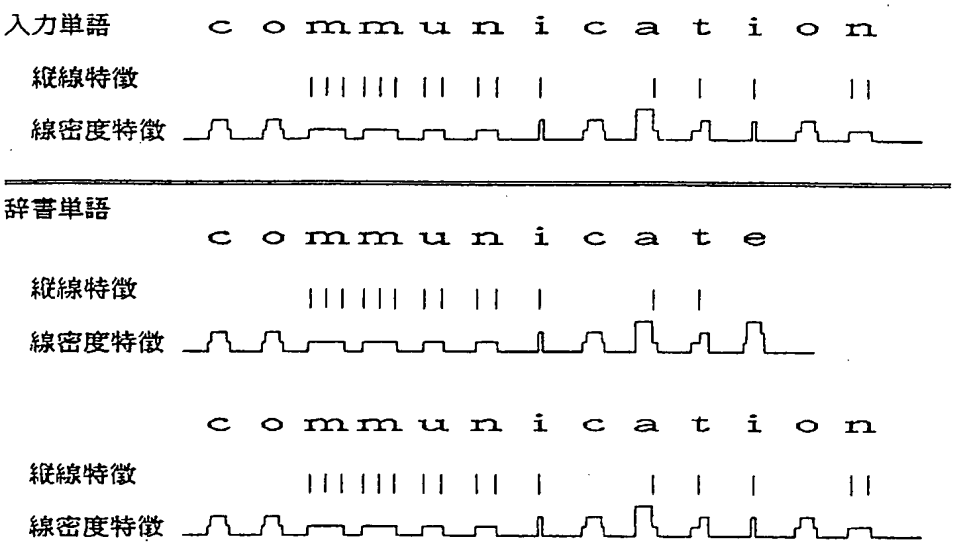
本発明の原理説明図

第１図



本発明の実施例構成図

第2図



本発明の認識処理説明図

第3図